College major on Income

Bilal Ahmad

11/09/2021

Introduction

We'll look at a College data and my assignment is to study how income varies across college major categories.

A codebook for the dataset is given below:

- rank: Rank by median earnings
- major_code: Major code
- major: Major description
- major_category: Category of major
- total: Total number of people with major
- sample_size: Sample size of full-time, year-round individuals used for income/earnings estimates: p25th, median, p75th
- p25th: 25th percentile of earnings
- median: Median earnings of full-time, year-round workers
- p75th: 75th percentile of earnings
- perc_men: % men with major (out of total)
- perc_women: % women with major (out of total)
- perc_employed: % employed (out of total)
- perc_employed_fulltime: % employed 35 hours or more (out of employed)
- perc_employed_parttime: % employed less than 35 hours (out of employed)
- perc_employed_fulltime_yearround: % employed at least 50 weeks and at least 35 hours (out of employed and full-time)
- perc_unemployed: % unemployed (out of employed)
- perc_college_jobs: % with job requiring a college degree (out of employed)
- perc_non_college_jobs: % with job not requiring a college degree (out of employed)
- perc_low_wage_jobs: % in low-wage service jobs (out of total)

The specifically question for this project is: "Is there an association between college major category and income?"

Based on your analysis, would you conclude that there is a significant association between college major category and income?

Load data

```
library(collegeIncome)

data(college)
```

Some exploratory analysis

```
head(college,5)
```

```
##   rank major_code                                    major major_category
## 1    1       2419                      Petroleum Engineering    Engineering
## 2    2       2416             Mining And Mineral Engineering    Engineering
## 3    3       2415                  Metallurgical Engineering    Engineering
## 4    4       2417 Naval Architecture And Marine Engineering    Engineering
## 5    5       2405                       Chemical Engineering    Engineering
##   total sample_size perc_women p25th median  p75th   perc_men
perc_employed
## 1  2339          36  0.9109326 25000  40000  50000 0.08906743
0.9115044
## 2   756           7  0.5154064 26000  37000  40000 0.48459355
0.7980501
## 3   856           3  0.5942076 26700  45000  60000 0.40579235
0.7871943
## 4  1258          16  0.6521298 26000  35000  45000 0.34787018
0.8465608
## 5 32260         289  0.4179248 31500  62000 109000 0.58207520
0.8515625
##   perc_employed_fulltime perc_employed_parttime
## 1              0.9206524              0.1774785
## 2              0.7110092              0.3623853
## 3              0.8833498              0.3387257
## 4              0.9366337              0.1673267
## 5              0.8086363              0.4020061
##   perc_employed_fulltime_yearround perc_unemployed perc_college_jobs
## 1                        0.7704431      0.08849558         0.6702970
## 2                        0.7093101      0.20194986         0.3867764
## 3                        0.7738366      0.21280567         0.7289116
## 4                        0.6527853      0.15343915         0.2460902
## 5                        0.6852821      0.14843750         0.5867515
##   perc_non_college_jobs perc_low_wage_jobs
## 1             0.1821782         0.05544554
## 2             0.5158761         0.21560172
```

```
## 3             0.1759983         0.03014828
## 4             0.4107636         0.04323827
## 5             0.3860437         0.11801062
```

```
str(college)
```

```
## 'data.frame':    173 obs. of  19 variables:
##  $ rank                          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ major_code                    : int  2419 2416 2415 2417 2405 2418
6202 5001 2414 2408 ...
##  $ major                         : chr  "Petroleum Engineering" "Mining
And Mineral Engineering" "Metallurgical Engineering" "Naval Architecture And
Marine Engineering" ...
##  $ major_category                : chr  "Engineering" "Engineering"
"Engineering" "Engineering" ...
##  $ total                         : int  2339 756 856 1258 32260 2573
3777 1792 91227 81527 ...
##  $ sample_size                   : int  36 7 3 16 289 17 51 10 1029 631
...
##  $ perc_women                    : num  0.911 0.515 0.594 0.652 0.418
...
##  $ p25th                         : num  25000 26000 26700 26000 31500
23000 32500 37900 29200 23000 ...
##  $ median                        : num  40000 37000 45000 35000 62000
44700 45000 57000 36000 32200 ...
##  $ p75th                         : num  50000 40000 60000 45000 109000
50000 58000 67000 46000 47100 ...
##  $ perc_men                      : num  0.0891 0.4846 0.4058 0.3479
0.5821 ...
##  $ perc_employed                 : num  0.912 0.798 0.787 0.847 0.852
...
##  $ perc_employed_fulltime        : num  0.921 0.711 0.883 0.937 0.809
...
##  $ perc_employed_parttime        : num  0.177 0.362 0.339 0.167 0.402
...
##  $ perc_employed_fulltime_yearround: num  0.77 0.709 0.774 0.653 0.685 ...
##  $ perc_unemployed               : num  0.0885 0.2019 0.2128 0.1534
0.1484 ...
##  $ perc_college_jobs             : num  0.67 0.387 0.729 0.246 0.587 ...
##  $ perc_non_college_jobs         : num  0.182 0.516 0.176 0.411 0.386
...
##  $ perc_low_wage_jobs            : num  0.0554 0.2156 0.0301 0.0432
0.118 ...
```

We can see that the data has 173 observations of 19 variables which corresponds to the codebook. The question asks about relationship between the major category and income, so I will only look at major_category and median. There are obviously other factors that may affect our analysis, for example: gender perc_men and perc_women, sample size (number of objects that provide income) perc_employed and total… I assume to omit all other variables.

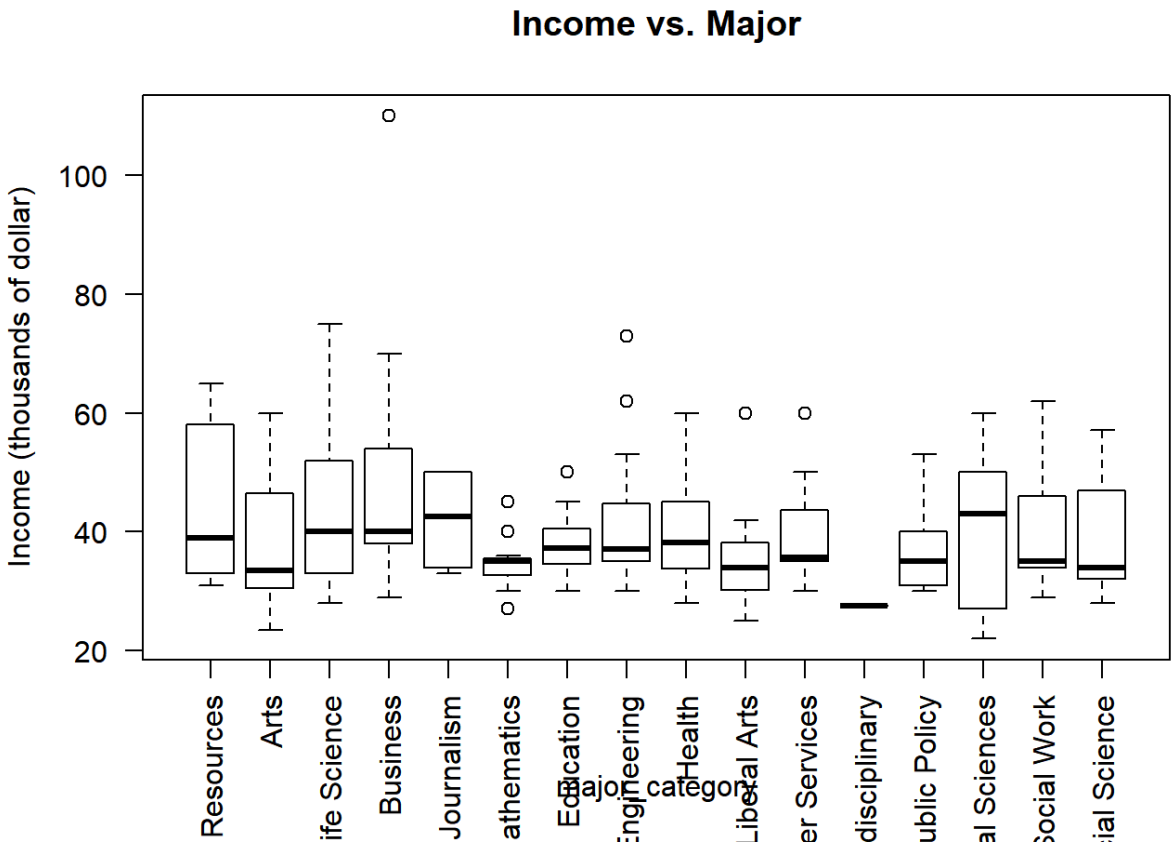
Now let's factorize the data and see the relationship between our two interested values:

```
college$major <- as.factor(college$major)

college$major_code <- as.factor(college$major_code)

college$major_category <- as.factor(college$major_category)


boxplot(median/1000 ~ major_category, data = college, main = "Income vs.
Major", ylab="Income (thousands of dollar)", las = 2)
```



We can see the distribution of the median of Income of each major is not normal, they're skewed. However for the purpose of this project of practicing linear model, I assume they're normal.

Analyze

Let's have a look at the major_category:

```
unique(college$major_category)
```

```
##  [1] Engineering                          Business
##  [3] Physical Sciences                    Law & Public Policy
##  [5] Computers & Mathematics              Agriculture & Natural Resources
##  [7] Industrial Arts & Consumer Services  Arts
##  [9] Health                               Social Science
## [11] Biology & Life Science               Education
## [13] Humanities & Liberal Arts            Psychology & Social Work
## [15] Communications & Journalism          Interdisciplinary
## 16 Levels: Agriculture & Natural Resources Arts ... Social Science
```

There are 16 of them. Let's first reorder the category before doing regression model:

```
college <- college[order(college$major_category),]
```

When we apply a linear model to this data, linking Income to all Majors, the default output intercept is the mean of the referenced major (alphabet sorted, with Agriculture first), the gradient coefficient of other majors is the difference of the mean of that major to the referenced one, and the p-value of those coefficients is the probability of a t-test if that mean and the referenced mean is different. For example, say we want to compare major Arts with others:

```
major_category_ref <- relevel(college$major_category, "Arts")
fit <- lm(median ~ major_category_ref, data = college)
summary(fit)$coef
```

```
##                                                    Estimate Std.
Error
## (Intercept)                                       38050.000
4014.658
## major_category_refAgriculture & Natural Resources  5450.000
5386.228
## major_category_refBiology & Life Science           5814.286
5032.640
## major_category_refBusiness                        11103.846
5102.541
## major_category_refCommunications & Journalism      3950.000
6953.591
## major_category_refComputers & Mathematics         -3331.818
5276.294
## major_category_refEducation                        -112.500
4916.931
```

```
## major_category_refEngineering                                2343.103
4534.719

## major_category_refHealth                                     2266.667
5182.901

## major_category_refHumanities & Liberal Arts                 -2883.333
4971.264

## major_category_refIndustrial Arts & Consumer Services        2378.571
5876.857

## major_category_refInterdisciplinary                        -10550.000
12043.973

## major_category_refLaw & Public Policy                        -250.000
6473.441

## major_category_refPhysical Sciences                          2350.000
5386.228

## major_category_refPsychology & Social Work                   1838.889
5517.619

## major_category_refSocial Science                             1016.667
5517.619

##                                                               t value
Pr(>|t|)

## (Intercept)                                                 9.47776950
3.919976e-17

## major_category_refAgriculture & Natural Resources           1.01183974
3.131715e-01

## major_category_refBiology & Life Science                    1.15531531
2.497166e-01

## major_category_refBusiness                                  2.17614057
3.103954e-02

## major_category_refCommunications & Journalism               0.56805181
5.708113e-01

## major_category_refComputers & Mathematics                  -0.63146941
5.286520e-01

## major_category_refEducation                                -0.02288012
9.817749e-01

## major_category_refEngineering                               0.51670312
6.060905e-01

## major_category_refHealth                                    0.43733553
6.624690e-01

## major_category_refHumanities & Liberal Arts                -0.58000007
5.627460e-01

## major_category_refIndustrial Arts & Consumer Services       0.40473529
6.862230e-01
```

```
## major_category_refInterdisciplinary                    -0.87595680
3.823917e-01

## major_category_refLaw & Public Policy                  -0.03861934
9.692429e-01

## major_category_refPhysical Sciences                     0.43629787
6.632200e-01

## major_category_refPsychology & Social Work              0.33327579
7.393708e-01

## major_category_refSocial Science                        0.18425822
8.540487e-01
```

From this result we can get some information: - mean of median of Income from major Arts is 38,050 - difference of mean of median of Income of Agriculture & Natural Resources from Arts is 5,450, and p-value of this difference is 0.31, which implies that the difference is not significant - the same interpretation can be done for coefficients of other variables

For this project, we ideally run linear regression models of income (median) vs. college major (major_catecory) for all majors as referenced. Given a referenced level, the model coefficients will indicate the difference of the mean of other variables and the probability if they are the same. I will run regression model for each major as the reference. The similar probabilities are stored in a 2D matrix A.

```
A <- matrix(, nrow = 16, ncol = 16)


for (i in 1:16){
    major_category_ref <- relevel(college$major_category,
as.character(unique(college$major_category)[i]))
    fit <- lm(median ~ major_category_ref, data = college)
    tmp <- summary(fit)$coef[,4]
    # swap the first element to the corresponding position in the diagonal
matrix
    tmp1 <- tmp[1:i]
    tmp1 <- c(0,tmp1)
    tmp1 <- c(tmp1[-2],tmp1[2])
    tmp1 <- tmp1[-1]
    # save to A
    A[,i] <- c(tmp1,tmp[-(1:i)])
}
```

Edit the matrix and plot.

```
library(reshape)
```

```
library(ggplot2)
```

We should expect a square symmetric matrix, with diagonal values are very low.

```
B <- data.frame(A)
names(B) <- unique(college$major_category)
B$major <- unique(college$major_category)
Bmelt <- melt(B)
```